En busca del cisne negro: Análisis de la evidencia estadística sobre fraude electoral en Venezuela*

Ricardo Hausmann
Harvard University

Roberto Rigobon
Massachussets Institute of Technology

Septiembre 3, 2004

* Este estudio fue solicitado por Súmate quien además nos preparó las bases de datos que utilizamos. Agradecemos el muy importante esfuerzo de recolección de información que esto representó para dicha organización. Igualmente estamos muy en deuda con un asiduo colaborador quien por razones institucionales debe permanecer anónimo. Además agradecemos a Andrés Velasco por muy útiles comentarios. Las opiniones emitidas en este reporte y los errores que hayamos podido cometer son nuestra responsabilidad y no comprometen ni a Súmate ni a las instituciones universitarias a las que pertenecemos.

Abstract

Este estudio analiza diversas hipótesis de fraude electrónico en el Referéndum Revocatorio realizado en Venezuela el 15 de agosto de 2004. Definimos fraude como una diferencia entre la intención del elector y el registro oficial de los votos. Partimos de la hipótesis de que no hubo fraude e intentamos buscar evidencia que nos permita rechazar dicha hipótesis. Rechazamos la hipótesis de que se haya cometido un fraude mediante la aplicación de topes numéricos en las máquinas de algunos centros. Igualmente, rechazamos cualquier hipótesis que implique alterar unas máquinas y no otras a nivel de cada centro electoral, pues los patrones de variación entre máquinas a nivel de cada centro de votación son normales. Sin embargo, la evidencia estadística es compatible con la ocurrencia de un fraude que haya afectado proporcionalmente a todas las máquinas de un mismo centro pero diferencialmente más a unos centros que a otros. Encontramos que el patrón de desviaciones entre centros de votación en la relación entre las firmas del Reafirmazo de Noviembre del 2003 y los votos por la opción del Sí el 15 de agosto votos está positiva y significativamente correlacionado con el patrón de desviaciones en la relación entre exit polls y votos. En otras palabras, en aquellos centros en que, de acuerdo al número de firmantes hay un inusualmente bajo número de votos Sí, es también donde de acuerdo a los exit polls ocurre lo mismo. Utilizando técnicas estadísticas descartamos que esto se deba a errores espúreos en la data o a coeficientes aleatorios en dichas relaciónes. Interpretamos que ello se debe a que tanto las firmas como los exit polls son medidas imperfectas de la intención del elector más no del posible fraude y por lo tanto lo que causa su correlación es precisamente la presencia de fraude. Además, encontramos que la muestra utilizada para la auditoría realizada el 18 de agosto no fue aleatoria ni representativa del universo de centros. En dicha muestra las firmas del Reafirmazo están asociadas a 10 por ciento más votos que en los centros no auditados. Construimos 1000 muestras aleatorias de los centros no auditados y encontramos que este resultado ocurre con una frecuencia inferior al 1 por ciento. Este resultado es compatible con la hipótesis de que la muestra para la auditoría fue escogida aleatoriamente pero solo entre aquellos centros cuyos resultados no habrían sido alterados.

Introducción

Este informe presenta los resultados de una evaluación estadística de los resultados del referendo revocatorio del mandato del Presidente Hugo Chávez Frías, convocado el 15 de agosto de 2004.

Desde la madrugada del 16 de agosto, en que el Consejo Nacional Electoral (CNE) anunció los resultados, voceros de la oposición expresaron dudas sobre la validez del resultado y argumentaron que se trataba de un fraude electrónico. Con el pasar del tiempo las dudas no se han despejado y la oposición sigue sin reconocer la victoria de la opción del NO en el referendo.

En este contexto, Súmate nos solicitó que realizáramos un análisis estadístico para verificar si la información disponible es compatible con una hipótesis de fraude o si, en cambio, rechaza dicha hipótesis. Súmate nos proporcionó los datos utilizados en este estudio pero nos dio completa autonomía sobre la forma de realizar nuestro trabajo así como sobre las conclusiones a las que llegamos.

Según fuimos informados, la presunción de fraude se basa en los siguientes elementos:

1- Se utilizó un nuevo sistema de votación automatizado, a pesar de que la oposición había pedido un conteo manual.

2- Las máquinas escogidas para el acto de votación emitían unas papeletas que le permitía a cada elector verificar que la máquina había contabilizado su voto adecuadamente. Sin embargo, el CNE decidió no permitir el conteo de las papeletas. En vez, decidió solamente hacer una “auditoria en caliente” sobre el 1 por ciento de las máquinas. Además, el CNE decidió que los números de las cajas a ser abiertas fueran escogidos por un programa generador de números aleatorios instalado en su propio computador.

3- Luego de una difícil negociación, el CNE permitió la participación de la OEA y el Centro Carter como observadores en todas las fases del proceso de votación salvo en la sala de totalización en la que el servidor central se comunicaba con los centros de votación. Tampoco se permitió la participación de ningún testigo de la oposición en dicha sala, cuyo acceso hasta el momento de totalización estuvo restringido a dos personas.

4- La tecnología adoptada permitía--de hecho requería--, la comunicación bi-direccional entre los servidores centrales y las máquinas de votación. Tal comunicación bi-direccional ocurrió.

5- Contrariamente a lo que estaba estipulado inicialmente, las máquinas de votación se comunicaron con sus respectivos servidores antes de imprimir las actas de totalización. Esto abre la posibilidad de que las máquinas fueran instruidas a

imprimir en las actas de votación un resultado distinto del expresado por los electores.

6- Durante el 15 de agosto, diversas organizaciones, incluyendo Súmate realizaron “exit polls” o encuestas a los votantes al salir estos de los CDVs. Para asegurar su calidad, la encuesta de Súmate fue realizada con la asistencia de la empresa Penn, Shoen and Berland. La misma arrojó un resultado radicalmente distinto al oficial. Lo mismo ocurrió con el exit poll realizado por Primero Justicia. La base de datos de ambas encuestas fue entregada a nosotros para la realización del presente estudio.

7- La auditoria en caliente realizada en la madrugada del 16 de agosto de 2004 no se cumplió a satisfacción de la oposición ni de los observadores internacionales, pues solamente llegaron a contarse 78 de las 192 cajas estipuladas. La oposición solamente asistió a 28 de estos conteos, y los observadores internacionales solamente estuvieron presentes en menos de 20 de estos conteos.

8- A solicitud de los observadores internacionales se realizó una segunda auditoría el 18 de agosto. Esta no gozó de la participación de la oposición pues no se cumplieron las condiciones que habían exigido, tales como el transporte del material electoral a un lugar central antes de ser escogidas las cajas a ser abiertas y la verificación de que éstas fueran las originales, constatando que en las papeletas se encontraran las huellas digitales de los votantes. En cambio, las cajas se escogieron 24 horas antes de ser abiertas, lo que en teoría daría tiempo para que fueran alteradas. Además, el CNE insistió en que no se utilizara el programa de generación de números aleatorios propuesto por el Centro Carter, sino que en vez se usara su propio programa. Esto plantea dudas sobre si la muestra seleccionada es realmente aleatoria.

Todos estos hechos abren la posibilidad de que haya habido un fraude electrónico en el cual las máquinas imprimieron actas con un resultado distinto al real. Esto pudo haber sido realizado ya sea mediante alteraciones del software o mediante la comunicación electrónica con el centro de cómputos.

En este contexto, Súmate nos pidió que realizáramos un estudio estadístico para establecer en que medida los datos son compatibles con estas hipótesis de fraude. El problema consiste en encontrar una metodología capaz de identificar manipulaciones de la voluntad del elector que pudieran dejar huellas identificables en la estructura de los votos y en su relación con otras bases de información.

Análisis de los exit polls

La primera evidencia de potenciales irregularidades en el cómputo proviene de los exit polls realizados independientemente por Súmate y por Primero Justicia (PJ). Como muestra el Cuadro 1, de acuerdo al CNE la votación obtenida por el Sí fue del 41.1 por

ciento. En cambio, el las encuestas de Súmate y de PJ la proyección ponderada fue de 62.0 y de 61.6 por ciento respectivamente, una diferencia de más de 20 puntos.

En principio, esta diferencia puede deberse a que la muestra escogida por Súmate y Primero Justicia no era representativa del universo electoral. Es decir, es posible que en las muestras de centros escogidos haya una sobre-representación de centros a favor del Si con respecto a los que están a favor del No. Sin embargo, ésta no una parte importante de la explicación de la diferencia. Como muestra el Cuadro 1, según el CNE el porcentaje obtenido por el Sí en los centros encuestados por Súmate fue del 45.0 por ciento, mientras que en la muestra del PJ el resultado fue de 42.7 por ciento. Es decir, en la muestra escogida por ambas organizaciones, el resultado reportado por ellas difiere del oficial en mas de 17 puntos porcentuales. Por ello, la diferencia en los resultados no se debió a la composición de la muestra sino más bien a una diferencia sistemática en el promedio de los centros de votación auditados.

Cuadro 1. Comparación entre los reultados electorales y los exit polls de Súmate y Primero Justicia

	Unweighted	Weighted
Porcentaje del Sí en votos a nivel de centro	37.0%	41.1%
Porcentaje del Sí en el exit poll de Súmate	59.5%	62.0%
Porcentaje del Sí en los votos de los centros donde Súmate realizó su exit poll	42.9%	45.0%
Porcentaje del Sí en el exit poll de PJ	62.6%	61.6%
Porcentaje del Sí en los votos de los centros donde PJ realizó su exit poll	42.9%	42.7%
Porcentaje del Sí en los exit polls de Sumate+PJ	61.3%	62.2%
Porcentaje del Sí en los votos de los centros donde Súmate+PJ realizaron su exit poll	43.10	44.2%

Para ilustrar este problema más claramente, en el Gráfico 1 mostramos el porcentaje de votos y los resultados de las encuestas para los 340 centros encuestados por ambos grupos. Si las encuestas fuera perfectas, los puntos se alinearían en una línea recta de 45 grados que pasa por el origen (dibujada en el gráfico). Es decir, donde los votos por el Si son 10 por ciento las encuestas darían el mismo resultado. Esto mismo ocurriría con los centros que arrojan el 20, el 50, el 80 o el 100 por ciento de votos. Si los puntos del gráfico están por encima de la curva de 45 grados quiere decir que en general la encuesta sobre-estima el resultado, centro por centro. Si los puntos están por debajo, las encuestas lo subestiman.

Como el grafico muestra claramente, de los 340 centros encuestados, la gran mayoría está por encima de la recta de 45 grados. Además, el gráfico indica que las diferencias entre los votos y las encuestas son muy variables entre centros. Las distancias con la recta de 45 grados son mayores en los lugares donde el Si sacó entre 20 y 40 por ciento.

Este análisis tiene las siguientes implicaciones. Primero, indica que la diferencia entre las encuestas y los votos no se debe en forma importante a problemas de la representatividad de la muestra. Segundo, el análisis implica que la diferencia puede deberse a una de dos razones o a una combinación de ambas. Puede deberse a una falla generalizada de ambas encuestas en cada uno de los centros de votación, o a una manipulación bastante generalizada y no lineal de los resultados. Será parte del reto del trabajo estadístico distinguir estas dos hipótesis e investigar cuál es la correcta.

Gráfico 1. Exit polls vs. Resultado electoral: porcentaje del Si por Centro de Votación



La hipótesis de los topes

La hipótesis de fraude más discutida en Venezuela ha estado basada en la idea de que se le impuso topes numéricos a la cantidad de votos Si que podían sacarse en un centro y que la diferencia se habría asignado al No. En esta sección evaluamos esta hipótesis.

Para analizar la factibilidad de esta hipótesis miramos en la base de datos del CNE que contiene 19062 cuadernos automatizados cuántas veces el número de votos Sí y de votos No se repite a nivel de Centro de Votación (CDV).

Cuadro 2. Número de votos Si y No que se repiten en los cuadernos de un mismo centro

Variable	Num. Cuadernos	Num Repetidos	Frecuencia
Si	19062	1875	9.8
No	19062	1472	7.7

La repetición del Si ocurre con una frecuencia de 9.8 por ciento mientras que las del No ocurre con una frecuencia del 7.7 por ciento. La frecuencia relativamente alta es explicable por el hecho de que tanto el número de electores como el porcentaje de votación tienden a parecerse mucho entre los cuadernos de un mismo centro. El primero, porque así se diseñan los cuadernos y el segundo porque en cada centro, los votantes se distribuyen aleatoriamente por cuaderno. El hecho de que el Si ocurra con una frecuencia ligeramente más alta que el No es explicable porque el Si tiene un porcentaje de votos más bajo. Ilustremos este punto con un ejemplo. Supongamos que la preferencia por el Si en un centro es aproximadamente del 40 por ciento y que el número de votantes en cada cuaderno es 100. Una variación porcentual del 5 por ciento implicaría 2 votos, por lo que el resultado esperado en cada máquina estaría entre 38 y 42. El resultado pudiera estar en algunos de los 5 números incluídos en ese intervalo. En cambio, la misma variación porcentual para el No daría una variación entre 57 y 63 votos, lo cual da 7 números posibles. Como la cantidad de números posibles es más alta para el No que para el Si, es lógico que éste se repita con menos frecuencia.

Además, la hipótesis de los topes implica que el número que se repite es también el máximo del centro y que la diferencia se le asigna al No. Para ello, es necesario que el número repetido sea también el máximo del centro de votación. Estudiamos esta hipótesis en el Cuadro 3.

Si el número repetido estuviese aleatoriamente distribuído, ocurriría con una frecuencia igual a 1/(Número de cuadernos – 1). Por ejemplo, en el caso de 2 cuadernos ambos números repetidos son simultáneamente el máximo y el mínimo, pues hay un solo número. En el caso de tres cuadernos, la probabilidad de que el que se repita sea el máximo es del 50 por ciento. Como vemos en el Cuadro 3, 66 no está muy lejos de ser la mitad de 124. En el caso de 5 cuadernos, 54 no está lejos de ser la cuarta parte de 198.

Concluimos que si hubo fraude, éste no se realizó por vía de la imposición de topes numéricos a los votos Si en las máquinas de un centro de votación.

Cuadro 3. Números repetidos máximos y no máximos por número de cuadernos en los centro de votación

Cuadernos Por centro	No Máximos	Máximos	Total
2	0	64	64
3	58	66	124
4	161	80	241
5	144	54	198
6	230	46	276
7	221	46	267
8	197	14	211
9	151	4	155
10	97	8	105
11	85	2	87
12	52	2	54
13	36	0	36
14	18	0	18
15	20	0	20
16	7	0	7
17	6	0	6
18	6	0	6
Total	1,489	386	1,875

Análisis de la varianza de las máquinas de un mismo centro

La hipótesis de los topes, de ser cierta, también afectaría la diferencia en el porcentaje de votos que dan las máquinas de un mismo centro. Esto se debe a que la cantidad de votantes por mesa varía debido a diferencias en la tasa de abstención o en el número de electores en cada cuaderno electoral. Dicha variación se vería reflejada en el número de votos No y por tanto causaría una variación en el porcentaje que cada opción obtendría en las distintas máquinas. Esta hipótesis y cualquier hipótesis que se base en la idea de alterar unas máquinas más que otras a nivel del centro de votación (CDV) puede ser evaluada analizando la varianza de los resultados entre máquinas de un mismo centro.

A nivel de cada CDV, los votantes son distribuidos entre mesas y cuadernos de acuerdo a los dos últimos dígitos del número de su cédula de identidad. Ello hace que cada cuaderno sea una muestra aleatoria de los votantes del centro, pues los últimos dígitos de la cédula de identidad no están correlacionados con ninguna variable relevante a la decisión de votación. Esto limita quétan distintos pueden ser los resultados de dos máquinas del mismo centro. Para ilustrar esto, pensemos primero en cómo se hacen las encuestas de opinión en cualquier país. Se hace una muestra aleatoria –usualmente de 1000 o 2000 personas-- y con eso se intenta estimar el resultado para un país de 14 millones de electores. Es decir, la estimación se hace con algo así como 2 diezmilécimos de la población total. En el caso de un centro de votación, estamos tomando un universo mucho más pequeño y homogéneo que un país y estamos dividiendo la población aleatoriamente de acuerdo al número de cuadernos en el centro. Por ejemplo, en el caso de un centro de 5 cuadernos, cada máquina representa aproximadamente el 20 por ciento

de la población total del centro. Además, en el caso de esta votación, las opciones se limitaban a dos: Sí o No. Esta estructura impone una condición para el la desviación típica del número de votos por máquina. Supongamos que en una máquina voten N personas y que la probabilidad de que cada una de ellas vote por el Si sea p. La teoría de probabilidades exige que la desviación típica sea igual a:

$$Desv.Tipica = \sqrt{p(1-p)N}$$

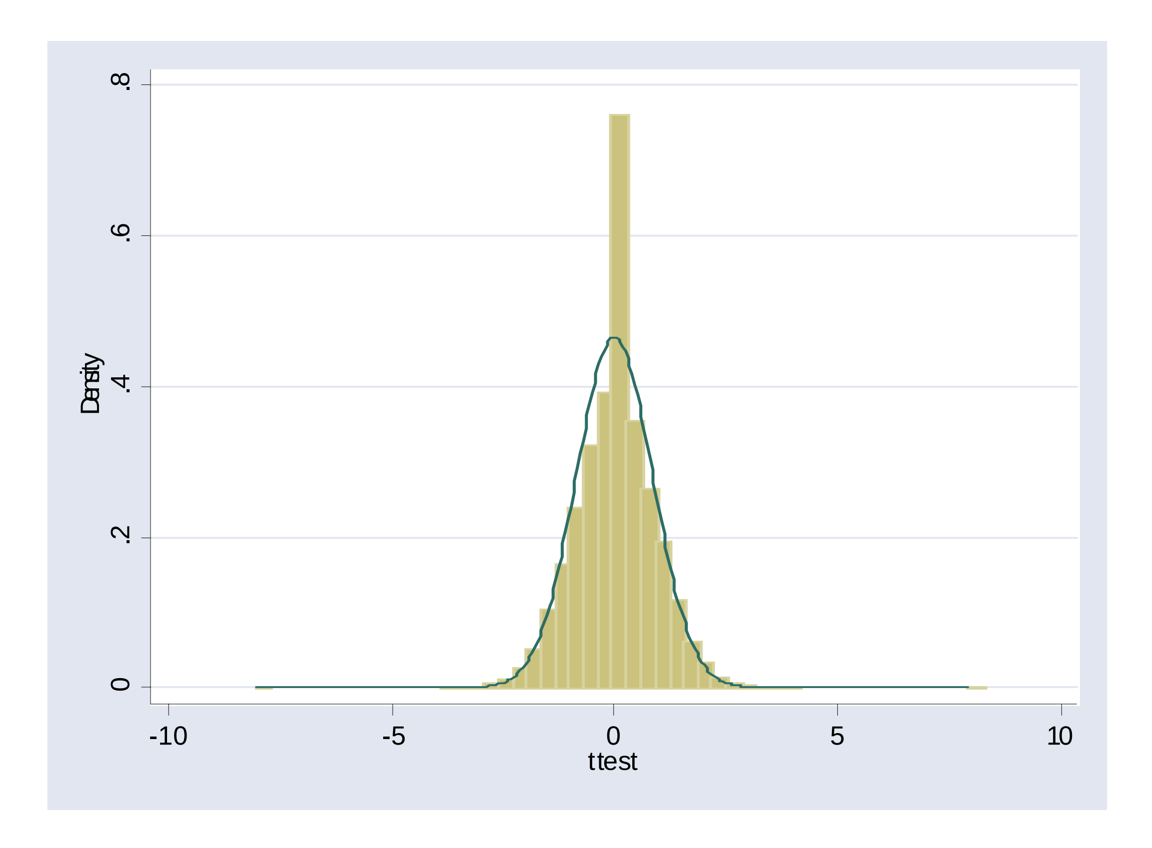
Para ilustrar esto, tomemos el caso en el que *p*, la probabilidad de que un elector vote por el Sí en un centro dado, es igual a 50 por ciento y que N sea 400. En este caso, la desviación típica será de 10 votos. La desviación típica será de 10 dividido por 400, es decir, del 2.5 por ciento. Dado esto, la desviación típica entre máquinas de un mismo centro debe ser compatible con esta regla.

Si por ejemplo, dentro de un mismo centro se cambiaran los resultados de unas máquinas en 10 por ciento mientras que las otras se dejaran inalteradas, entonces veríamos un aumento en la desviación entre máquinas que sería incompatible con la ley de la distribución binomial. Y, como dijimos anteriormente, si se le fijaran topes numéricos a todas las máquinas de un centro, las variaciones del número de votantes por cuaderno afectaría el número de votos del No y por tanto alteraría el porcentaje de votos SI de manera distinta en cada máquina, por lo que veríamos una desviación excesiva a nivel de centro.

Para verificar si la data de votos del CNE cumple con la desviación típica que predice la teoría de las probabilidades, calculamos la desviación de cada máquina con respecto al promedio porcentual de su centro de votación. Además, dividimos este número por la desviación típica que le correspondería a un centro de con ese número de votantes y de máquinas. El gráfico 2 sintetiza nuestros resultados. Muestra un histograma de frecuencias de la diferencia porcentual del SI entre máquinas de un mismo centro con respecto a la desviación típica que deberíamos observar. La curva refleja la distribución teórica. Las barras son las frecuencias calculadas en la data del CNE. Como se verá, hay una coincidencia muy grande entre la curva y las barras. El gráfico indica que sólo alrededor del 1 por ciento de los cuadernos tienen una desviación mayor que 2 veces la desviación típica que les corresponde, tal como predice la teoría. De hecho, si algo sorprende del gráfico es la gran concentración de resultados alrededor de variaciones porcentuales cercanas al cero.

Este resultado tiene dos interpretaciones posibles. Una es que no hubo fraude. La otra es que si el fraude se cometió, éste debió haberse hecho cambiando todas las máquinas de un mismo centro en un porcentaje similar. De hecho, un fraude de este tipo no sería detectado con el análisis hecho hasta el momento pues no alteraría la varianza de los resultados entre máquinas. Cualquier hipótesis de fraude que no cumpla con esta condición violaría la regla de la varianza de la distribución binomial.

Gráfico 2. Distribución de las desviaciones típicas porcentuales por cuadernos de cada centro de votación con respecto a su desviación teórica.



Una estrategia para detectar estadísticamente la presencia de fraude

Para detectar si los datos existentes son compatibles con la existencia de fraude necesitamos desarrollar un método estadístico adecuado. Definimos fraude como la diferencia entre la intención del votante y lo que registra el sistema electoral como resultado de su decisión.

(1) Votos = Intención – Fraude = I - F

El problema es que no podemos observar directamente la intención del elector. La estrategia estadística que adoptamos parte por encontrar dos conjuntos de variables independientes que estén correlacionadas con la intención de voto, más no así con el fraude. Para nuestros efectos, no es demasiado importante que nuestras variables no predigan la intención de voto perfectamente. Aunque lo hagan imperfectamente y con aleatoriedad, nos permiten en principio rechazar o no la hipótesis de que hubo fraude.

Para ilustrar lo que hacemos, conviene realizar una primera explicación simplificada del método adoptado. En la práctica, hacemos las cosas con un nivel más grande de

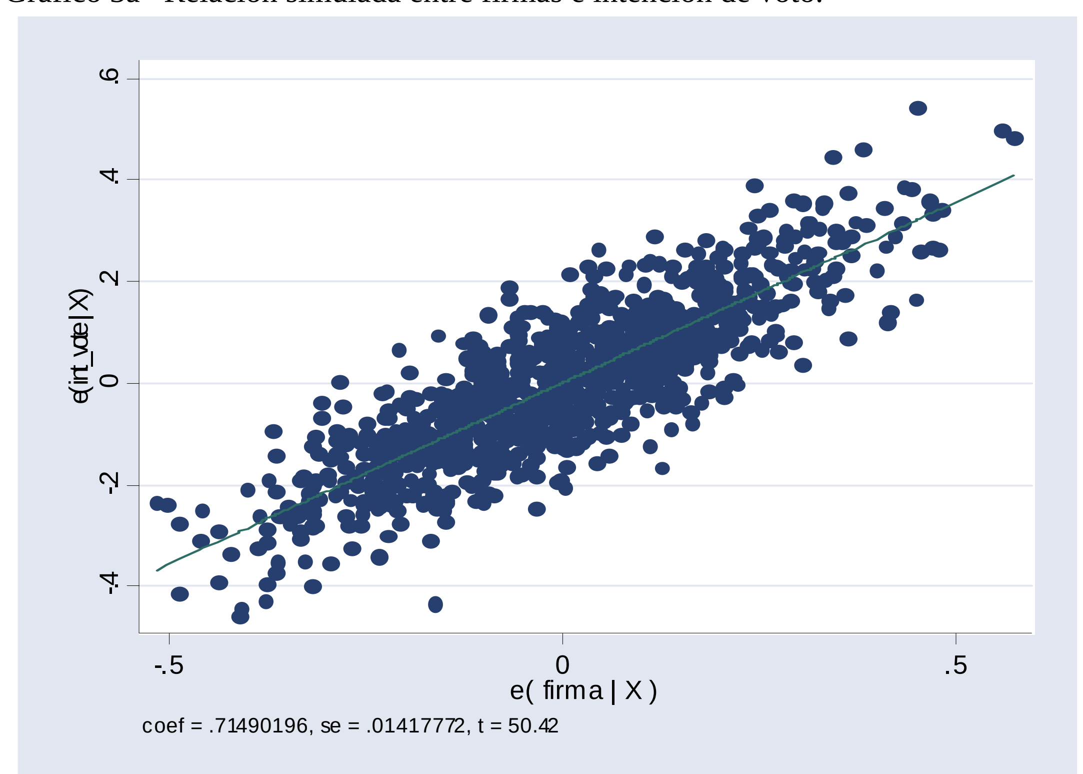
complejidad, pero este será más fácil de comprender si lo ilustramos primero con un caso simplificado.

Tomemos dos variables que están correlacionadas con la intención del elector: las firmas del reafirmazo realizado en Noviembre-Diciembre del 2003 y los exit polls. Cada una de estas variables es una medida imperfecta de lo que pudo haber sido la intención del elector el 15 de agosto del 2003. Alguna gente que firmó pudo haber cambiado de opinión. Otros decidieron no firmar pues se trataba de un acto público, pero si estaban dispuestos a votar Sí en agosto pues el voto es secreto. Otros no estaban inscritos en el Registro Electoral Permanente (REP) para noviembre, pero si lo estuvieron para agosto. Las colas en la elección de agosto fueron particularmente largas y lentas y eso pudo haber limitado la capacidad de algunas personas de expresarse electoralmente, etc.

Igualmente, los exit polls son una medida imperfecta de la intención de voto. Los encuestadores pudieron haber consciente o inconscientemente escogido una muestra sesgada. La gente pudo haber tenido más o menos intención de cooperar con la entrevista, etc. Lo importante es que ambas medidas deben estar correlacionadas con la intención del votante más no así con el fraude.

Supongamos por un instante que tuviésemos una medida perfecta de la intención de votos del elector en cada centro de votación y construyésemos un gráfico de nuestra variable, por ejemplo las firmas, y de dicha intención. Como las firmas son una medida imperfecta de la intención de votos, el gráfico se verá como una nube de puntos (gráfico 3a). Este gráfico fue construido con datos simulados usando un generador de números aleatorios. Los datos fueron creados suponiendo que cada firma genera 0.7 votos con un error normalmente distribuído entre + 0.1 y -0.1. Con una técnica denominada análisis de regresión podemos derivar la línea que relaciona las firmas con la intención de votos. La relación real es 1, pues así construimos los datos. Con el análisis de regresión lo que podemos calcular es que la relación estimada es 0.71 más o menos un intervalo de confianza de 0.014, como lo indica el gráfico.

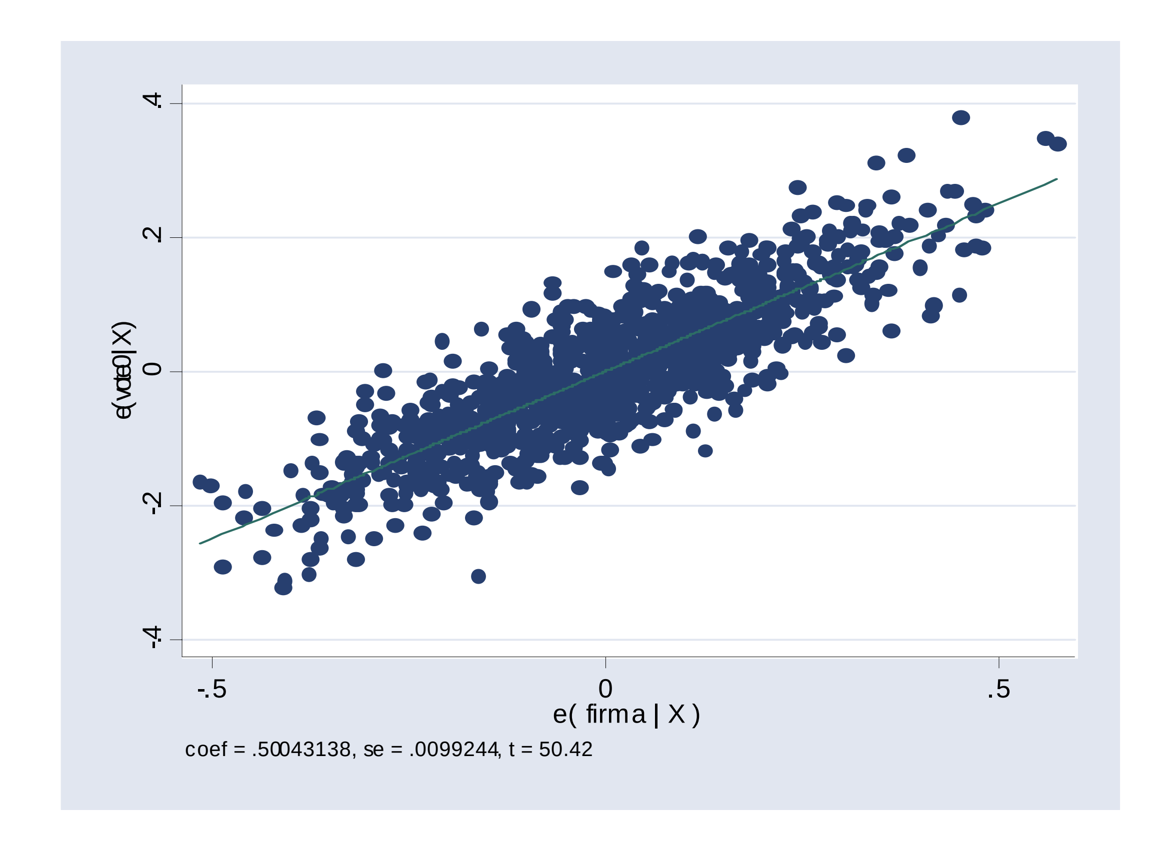
Gráfico 3a   Relación simulada entre firmas e intención de voto.



Ahora bien, nosotros no observamos la verdadera intención del elector sino los votos registrados y estos, en teoría, pueden estar influidos por el fraude. Supongamos que se realiza un fraude que es directamente proporcional al número de votos en ese centro. Por ejemplo, supongamos que el fraude implica multiplicar el total de votos Si en una máquina por un número menor que 1, digamos 0.7. Esto es equivalente a quitarle 30 por ciento de votos Sí. El gráfico 3b muestra las consecuencias de esto.

Ahora la pendiente estimada ya no es 0.7 sino 0.5, pero el patrón de errores --es decir, la distancia con respecto a la línea marcada--, no muestra ninguna anomalía. Si así fuese como se cometió el fraude, no podríamos detectarlo con nuestro método. De hecho, un fraude que rebaje un porcentaje fijo a todas las máquinas por igual sería prácticamente imposible de detectar por métodos puramente estadísticos. Se requeriría de una auditoría de las papeletas u otra fuente información.

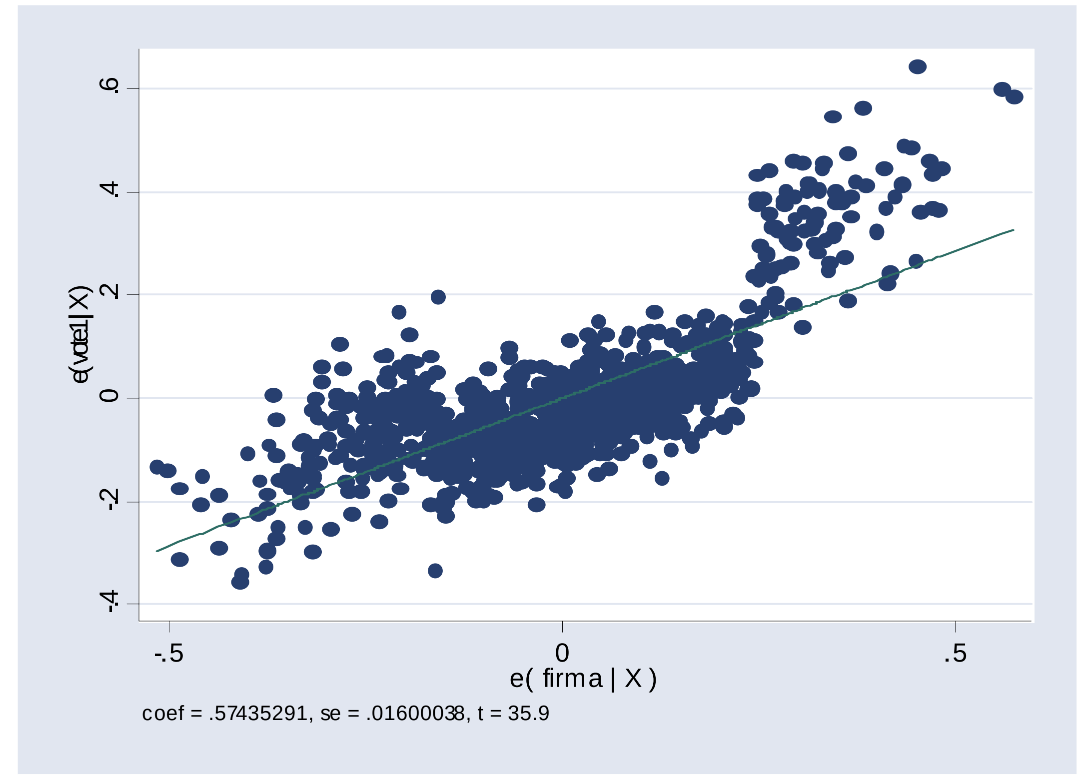
Gráfico 3b Relación simulada entre firmas y votos con fraude proporcional del 30 por ciento de los votos Sí.



Supóngase entonces que el fraude no fue realizado de manera directamente proporcional a los votos sino que se hizo en algunos centros y no en otros, o que se hizo más en unos que en otros. En particular, suponemos para ilustrar el caso, que el fraude consiste en eliminar el 30 por ciento de los votos Sí en centros en los que las firmas fueron inferiores al 30 por ciento de los electores o superiores al 70 por ciento. En ese caso, al observar las desviaciones entre la línea de regresión y los puntos en el gráfico se ve una anomalía. A la hora de predecir los votos los errores que cometen las firmas se debe no sólo a su imperfección sino también al fraude. (gráfico 3c).



Gráfico 3c. Regresión con fraude no proporcional: resta 30 por ciento cuando las firmas son inferiores a 0.3 o superiores a 0.7.



¿Que pasa si utilizamos ahora una segunda medida de la intención del elector, por ejemplo los exit polls? Esta también es una medida imperfecta de la intención de voto y por lo tanto al hacer un análisis de regresión, este va a generar algunos errores. Sin embargo, si hay un fraude no proporcional, este también va a generar una anomalía en el término de error, es decir en la desviación entre el dato real y la línea de regresión.

Nótese que cada medida – firmas y exit polls – son imperfectas. Sin embargo, lo que hace que cada una de ellas sea imperfecta son factores distintos e independientes entre si. El exit poll no es afectado por la tasa de abstención, pues se entrevista a la gente después de votar. Las firmas no dependen de la pericia o sesgos del entrevistador. La gente pudo haber cambiado de opinión entre noviembre y agosto, pero no hay la misma razón para que la gente cambie de opinión entre el acto de votación y la entrevista a la salida. La firma es un acto público y el voto es secreto, etc. Por lo tanto, los errores que comete cada medida pueden ser más o menos grandes pero no deben estar correlacionados. Sin embargo, si hay un fraude no proporcional, este afectará de la misma manera a cada una de las medidas. Por ello, los errores que éstas cometen debieran estar positivamente correlacionados.

Esta es la esencia del procedimiento que utilizamos. Corremos una regresión entre votos y firmas (más otras variables que entraremos a describir próximamente. Es decir, calculamos la mejor línea que pasa por la nube de puntos ente votos y las variables explicativas que usamos en semejanza al gráfico 3. Luego recuperamos los errores que

comete dicha regresión o línea. Hacemos lo propio con la relación entre votos y exit polls y recuperamos esos errores. Luego, estudiamos si estos dos conjuntos de errores están positivamente relacionados[1].

En la práctica hacemos las cosas de manera un poco más compleja. Incluimos en el análisis otras variables que también afectan el número de votos. Estas son el número de nuevos votantes y la tasa de abstención electoral en cada centro de votación. Los nuevos votantes no pudieron participar en el Reafirmazo pues no estaban inscritos en el REP. Mientras más nuevos votantes haya, mayor número de votos debieran haber. Ahora bien, el porcentaje de votos Si pudiese aumentar o disminuir dependiendo de la diferencia en preferencias políticas de los nuevos votantes con respecto a los inscritos con anterioridad. La tasa de abstención obviamente disminuye el número de votos y, al igual que en el caso anterior, lo puede hacer de manera diferenciada entre la opción del Si o del No.

Además, debemos decidir qué tipo de línea usar en la regresión. Tenemos varias opciones. Podemos usar una línea recta, una relación geométrica o una relación entre porcentajes. Es decir podemos relacionar votos con firmas (lineal), el logaritmo de los votos con el logaritmo de las firmas (geométrica) o el porcentaje de votos Si con el porcentaje de firmas sobre electores. Por razones técnicas, preferimos la forma logarítmica[2], sin embargo realizamos el análisis de las tres maneras para ver si nuestros resultados dependen de la forma funcional que adoptamos.

Un ejemplo de las estimaciones realizadas utilizando la forma logarítmica se presenta en el Cuadro 4. La ecuación estimada es:

$$LSI = a + b * LFIRMA + c * elc_now + d\ VEL + error$$

donde LSI es el logaritmo del número de votos SI, LFIRMA es el logaritmo del número de firmas en cada centro, elc_now es el porcentaje de nuevos votantes VEL es el porcentaje de participación electoral, y donde a, b, c y d son parámetros a ser estimados. El cuadro 4 muestra los resultados de nuestra estimación para los 342 centros para los que también tenemos exit polls, utilizando como método el más convencional : los mínimos cuadrados.

---

[1] Se define como error o desviación la diferencia entre el valor de la línea de regresión en un cierto punto y el valor de la observación que le corresponde. Gráficamente, es la distancia vertical entre los puntos en los graficos 3a, 3b y 3c y la línea de regresión dibujada.

[2] Dado que los centros de votación son de tamaños muy distintos la forma lineal genera problemas de heterosqueedasticidad (es decir los errores absolutos tienden a ser mas grandes en los centros más grandes lo que implica que no están normalmente distribuidos).

Cuadro 4. Estimación de la ecuación entre votos y firmas, electores nuevos y participación electoral

Source	SS	df	MS		
				Number of obs	= 342
				F(3, 338)	= 3582.70
Model	185.800888	3	61.9336295	Prob > F	= 0.0000
Residual	5.84296339	338	.017286874	R-squared	= 0.9695
				Adj R-squared	= 0.9692
Total	191.643852	341	.56200543	Root MSE	= .13148

LSI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LFIRMA	.9942821	.0099034	100.40	0.000	.974802	1.013762
elc_now	.4604462	.0375	12.28	0.000	.3866834	.5342089
VEL	.3311808	.0813913	4.07	0.000	.1710835	.4912781
_cons	.3059669	.0782436	3.91	0.000	.1520611	.4598727

La estimación permite explicar el 97 por ciento de la varianza de los votos entre centros. Estima los parámetros a, b, c y d con gran precisión. En particular, a es la constante, estimada en 0.306. El parámetro b es la elasticidad entre firmas y votos y está estimado en casi 1 (en realidad es 0.994). Ello implica que si un centro tiene el doble de firmas que otro centro, obtiene en promedio el doble de votos. El parámetro c es la elasticidad de los votos Si ante variaciones del porcentaje de nuevos electores. Está estimado en 0.46, lo que significa que si en un centro aumentase el número de electores en 100 por ciento, los votos por el Si aumentarían en 46 por ciento. El parámetro d es la elasticidad del número de votos Si frente a cambios en la tasa de participación y está estimado en 0.306, lo que indica que un aumento de la tasa de participación electoral en 10 por ciento causaría un aumento del número de votos Si en 3.06 por ciento.

Esta ecuación no indica la relación real entre intención de votos y sus variables explicativas, sino entre estas últimas y los votos reconocidos por el CNE. Al igual que en el Gráfico 3b, la posible presencia de fraude afecta los coeficientes estimados, sesgando las pendientes hacia la baja, y en parte se encuentra en el término de error.

La segunda ecuación que estimamos es la relación entre votos y exit polls también para los 342 centros para los que tenemos datos de éstos. La ecuación que estimamos es:

LSI = f + g*lex_si + h*VEL + j + error

Donde lex_si es el número de votos Si que predice la encuesta para ese centro. Las letras f, g, h y j son parámetros mientras que LSI y VEL ya han sido definidos. Los resultados aparecen en el Cuadro 5.

Cuadro 5. Estimación de la relación entre votos y las encuestas de salida (exit polls)

```
      Source |       SS       df       MS              Number of obs =     342
-------------+------------------------------           F(  3,   338) =  526.51
       Model |  157.862978     3  52.6209927           Prob > F      =  0.0000
    Residual |  33.7808737   338  .099943413           R-squared     =  0.8237
-------------+------------------------------           Adj R-squared =  0.8222
       Total |  191.643852   341   .56200543           Root MSE      =  .31614

------------------------------------------------------------------------------
         LSI |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lex_si |   .9701892    .025357    38.26   0.000     .9203118    1.020067
     elc_now |  -.6612884   .0868377    -7.62   0.000    -.8320987    -.490478
         VEL |   .4244489   .1957766     2.17   0.031     .0393549    .8095429
       _cons |   .0722736   .2086177     0.35   0.729    -.3380789    .4826261
------------------------------------------------------------------------------
```

De nuevo, la ecuación explica gran parte de la varianza del logaritmo de los votos (82 por ciento). La elasticidad estimada de las intenciones de acuerdo a las encuestas con los votos es de 0.97.

Estas estimaciones también estarían sesgados a la baja por la presencia de fraude. Sin embargo, el término de error reflejaría en parte no sólo la imperfección de los instrumentos utilizados sino también la posible presencia de fraude.

La estrategia entonces es analizar la correlación entre los errores de ambas ecuaciones. Dicha correlación es de 24 por ciento, la cual es sorprendentemente alta. Esto nos impide rechazar la hipótesis de fraude. Dicho de otro modo, en aquellos sitios en los que las firmas se equivocan proporcionalmente más en el sentido de predecir más votos Si que los obtenidos, las encuestas de salida también sobre-estiman relativamente más los votos alcanzados. Dado que ambas medidas son independientes, la implicación es que lo que tienen en común es el fraude.

Cuadro 6 Análisis de la relación entre los errores de las ecuaciones usando mínimos cuadrados

Covarianza	$9.3 * 10^{-3}$
Desvío típico de la covarianza	$2.8 * 10^{-3}$
T-Student sobre la covarianza	4.1
Probabilidad distinto de cero	0.999
Correlación	0.24

Este es nuestro primer resultado consistente con la hipótesis de fraude. Formalmente, podemos decir que no podemos rechazar la hipótesis de que se realizó fraude. La presencia de esta correlación indica que hay algo en común entre los errores cometidos por el exit poll y los errores cometidos por las firmas y esto es consistente con una diferencia entre la intención de voto del elector y los votos registrados.

Sin embargo, es posible argumentar que la correlación que observamos pudiera estar generada por dos fuentes. Una es el hecho de que nuestras medidas de la intención del elector son muy ruidosas o imperfectas y los errores en dichas variables pudieran generar problemas. El segundo es que estamos suponiendo coeficientes fijos entre firmas y votos o entre exit polls y votos y estos coeficientes pudiesen ser aleatorios. Esto abre la posibilidad de que la correlación que estamos encontrando haya sido generada por otros factores y no por el fraude.

Para descartar esta posibilidad, aplicamos una técnica estadística denominada "Variables instrumentales". La idea es que tanto las firmas como el exit poll tienen errores o ruido. Sin embargo, este ruido es independiente el uno del otro. Lo que tienen en común las variables es que ambas están relacionadas con la intención del elector. La técnica parte por usar en la regresión no las firmas directamente sino aquel componente de éstas que está relacionado o en línea con los exit polls. En la jerga estadística diríamos que usamos a los exit polls como instrumento para corregir o limpiar los errores en las firmas antes de estudiar su correlación con los votos. Simétricamente, usamos las firmas para limpiar los exit polls antes de relacionarlos con los votos. Después de haber hecho estas dos regresiones con variables instrumentales, tomamos los errores de cada una de ellas y estudiamos su correlación. Si los errores están correlacionados positivamente no podemos rechazar la hipótesis de que hubo fraude. La justificación teórica de esta metodología para el análisis de fraude que estamos realizando es discutida en detalle en la versión técnica de este documento.

El Cuadro 7 presenta la misma ecuación que el Cuadro 4 pero esta vez utiliza el método de variables instrumentales, usando exit polls como instrumento.

Cuadro 7. Regresión entre votos y firmas con exit polls como variable instrumental

Instrumental variables (2SLS) regression

Source	SS	df	MS		
Model	185.741458	3	61.9138192	Number of obs =	342
Residual	5.90239422	338	.017462705	F(3, 338) =	3013.34
Total	191.643852	341	.56200543	Prob > F =	0.0000
				R-squared =	0.9692
				Adj R-squared =	0.9689
				Root MSE =	.13215

LSI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LFIRMA	1.012645	.0110631	91.53	0.000	.9908834	1.034406
elc_now	.4792798	.0380142	12.61	0.000	.4045055	.5540541
VEL	.3067645	.0820558	3.74	0.000	.1453602	.4681688
_cons	.1718993	.0861817	1.99	0.047	.0023791	.3414194

Instrumented: LFIRMA
Instruments: elc_now VEL lex_si

Nótese que ahora el coeficiente de firmas aumenta ligeramente: de 0.994 en la estimación del Cuadro 4 a 1.013 en el Cuadro 7. Esto es usual, pues la existencia de errores o ruido en los datos tiende a reducir los coeficientes estimados con el método del Cuadro 4. Al limpiar o reducir el problema de errores en los datos se obtienen usualmente coeficientes más elevados.

El Cuadro 8 re-estima la misma ecuación que el Cuadro 5 pero usando variables instrumentales. En esta oportunidad, el coeficiente del exit poll (lex_si) aumenta de 0.97 a 1.17. Esto es de esperarse pues los datos de los exit polls, dada su naturaleza, tienen más ruido que la data de firmas y por tanto el método del Cuadro 5 sesga el coeficiente más hacia la baja que en el caso de las firmas.

Cuadro 8. Regresión entre votos y exit polls usando firmas como variable instrumental

```
Instrumental                     variables                     (2SLS)                    regression

       Source   |         SS         df         MS                Number   of   obs   =       342
-------------+------------------------------                  F(      3,       338)     =     517.96
        Model   |    151.228444       3    50.4094815              Prob   >   F        =    0.0000
     Residual    |     40.4154074      338    .119572211                R-squared       =    0.7891
-------------+------------------------------                   Adj      R-squared      =     0.7872
        Total   |    191.643852      341     .56200543              Root   MSE         =    .34579

------------------------------------------------------------------------------
          LSI   |        Coef.     Std.   Err.        t      P>|t|       [95%   Conf.   Interval]
-------------+----------------------------------------------------------------
       lex_si    |      1.176787       .030827       38.17      0.000         1.11615       1.237424
      elc_now    |     -.6829967      .0949936       -7.19      0.000       -.8698498      -.4961437
          VEL    |      .1627794      .2148175        0.76      0.449       -.2597683       .5853271
        _cons    |     -1.523351       .250735       -6.08      0.000       -2.016549      -1.030153
------------------------------------------------------------------------------
Instrumented:                                                                              lex_si
Instruments:                         elc_now                      VEL                      LFIRMA
------------------------------------------------------------------------------
```

Al estudiar la relación entre los errores de las variables generados por estas dos ecuaciones obtenemos los datos presentados en el Cuadro 9. El análisis indica que aún después de usar el método de variables instrumentales para corregir problemas de errores en las variables y de coeficientes aleatorios, la correlación entre los errores generados usando firmas y aquellos generados usando los exit polls disminuye sólo de 0.24 a 0.17 y se mantiene muy significativamente distinto de cero.

Cuadro 9 Análisis de la relación entre los errores de las 2 ecuaciones usadas para estimar el número de votos: Mínimos cuadrados vs. Variable Instrumentales

Concepto	Método mínimos cuadrados	Método Variables Instr.
Covarianza	$9.3 * 10^{-3}$	$7.7 * 10^{-3}$
Desvío típico	$2.8 * 10^{-3}$	$2.5 * 10^{-3}$
Probabilidad distinto de cero	0.999	0.991
Correlación	0.24	0.17
T-Student sobre la covarianza	4.1	3.1

Nuestra hipótesis inicial era que si hubiese un fraude no perfectamente proporcional, este generaría un patrón de errores que causaría una correlación positiva entre ambas variables. Encontramos dicha correlación positiva, lo que nos permite rechazar la hipótesis de que no hubo fraude. Los exit polls, a pesar de sus imperfecciones, tienden a equivocarse más en aquellos mismos sitios donde las firmas también se equivocan. Al utilizar el método de variables instrumentales hemos descartado de que la correlación se deba a problemas de errores en las variables o a coeficientes aleatorios. Este es el tipo de huella que dejaría un fraude no perfectamente proporcional.

Nuestra estrategia ha consistido en utilizar dos fuentes de información relacionadas a la intención de voto del elector pero no al posible fraude. Si usamos estas fuentes o variables para estimar imperfectamente los votos, el residuo o término de error contendrá no sólo las imperfecciones de nuestras fuentes de información sino también un componente asociado al fraude. Nuestra interpretación es que como las imperfecciones son independientes unas de otras y los residuos están correlacionados, ello se debe a la presencia del factor que tienen en común, es decir, el fraude.

La auditoría

Cualquier hipótesis de fraude necesita explicar cómo ocurrieron las auditorías y qué tan compatibles son los resultados de las auditorías con la existencia de fraude. En particular, debe tomar en cuenta los resultados de la segunda auditoría, realizada a solicitud y con la participación plena de los principales observadores internacionales, es decir, el Centro Carter y la OEA. La auditoría se basó en abrir 150 cajas, escogidas en principio aleatoriamente, las cuales supuestamente contienen las papeletas originales chequeadas por los votantes y por lo tanto reflejan su verdadera intención de voto. Si estas cajas no fueron violadas y si son realmente una muestra aleatoria del universo al que corresponden, la auditoría sería una prueba muy significativa en contra de la hipótesis de fraude. Por ello, es importante preguntarse como se pudo hacer un fraude, dado que se realizó esta auditoría. Es bueno destacar que cualquier hipótesis de fraude que involucre cambiar cientos de cajas implica una conspiración mucho más grande en términos del número de personas involucradas, lo que la haría más vulnerable a la delación.

Una hipótesis es que el fraude no se realizó en todos los centros sino sólo en un sub-conjunto de ellos. Para ejemplificar, supongamos que de los 4580 centros automatizados en nuestra base de datos, se alteraron los resultados en 3000 centros y no en los demás. Supongamos además que los 1580 centros no alterados fueron escogidos aleatoriamente. Ello implica que representarían una muestra balanceada del país desde el punto de vista regional y social. Lo mismo sería cierto de los 3000 centros cuyos resultados habrían sido alterados. Una motivación para hacer esto sería que se sabía que se realizarían auditorías ex post y que debían dejar un número de centros sin alterar para poder realizar las auditorías.

Nótese que si se alteran unos centros y no otros, el fraude no es perfectamente proporcional, por lo que al realizar un análisis como el de la sección anterior se hubiese encontrado la correlación positiva de los errores, la cual obtuvimos.

Si la elección de los centros a no ser alterados se hizo de esta manera, esto crea una complicación importante pero también abre una gran oportunidad. La complicación es que la elección de las cajas a ser auditadas no podría ser realmente aleatoria. Es crítico que la elección se hiciese entre los 1580 centros sin modificaciones y no entre los 3000 centros alterados. Ello solo es posible si se tiene control sobre el programa de selección aleatoria de las cajas a ser auditadas. En este sentido cabe destacar que el CNE se negó a utilizar el programa de generación de números aleatorios propuesto por el Centro Carter e insistió que se utilizara el programa propuesto por ellos e instalado en su computador.

La oportunidad que genera esta forma de resolver el problema de la auditoría es que cualquier muestra que se tome sobre los 1580 centros no alterados será una muestra representativa del país en el sentido social o regional. Esto hace más difícil saber si la muestra escogida fue realmente aleatoria, pues se asemejará al país en todas las dimensiones usualmente asociadas con representatividad, tales como la regional o la social.

Para resolver este problema debemos desarrollar una metodología que nos permita examinar si efectivamente la muestra escogida para la auditoría del 18 de agosto es una muestra aleatoria. Para entender el problema más claramente, llamemos a los centros no alterados gordos y a los alterados, flacos. La muestra escogida para la auditoría debe ser una muestra de puros centros flacos, mientras que el resto de los centros son una mezcla de gordos y flacos. Si pudiésemos "pesar" los centros auditados, podríamos ver que en promedio son más gordos que los centros no auditados.

El problema es que necesitamos desarrollar una metodología que pueda examinar si efectivamente los centros de la autoría pesan como los otros o si por el contrario tienen una contextura estadísticamente distinta.

La forma que proponemos es la siguiente. Existe un teorema en estadística que dice que si una relación se aplica al conjunto completo, cualquier muestra aleatoria de éste debe tener las mismas propiedades. Si estimamos una relación para el universo de centros no auditados y estimamos otra para los centros auditados, la segunda no puede ser estadísticamente diferente de la primera. De otra forma, no sería una muestra aleatoria y representativa.

Para implementar esta estrategia utilizamos nuevamente nuestro modelo que relaciona firmas, tasa de participación electoral y nuevos votantes con el número de votos. Estimamos esta relación sobre el universo de 4580 centros y miramos los coeficientes obtenidos. Luego los estimamos separadamente entre los centros auditados y los no auditados y miramos si los coeficientes son estadísticamente distintos.

Para ver si los resultados son distintos y calcular la significación estadística de la diferencia, es útil estimar las ecuaciones de la siguiente forma:

Votos = a + b * (vector de variables explicativas) +
c * d * (vector de variables explicativas)

donde a, b y c son parámetros a ser estimados y d es una variable "dummy" que vale 1 si se trata de centros que fueron auditadas y 0 si se trata de centros que no fueron auditados. Las cajas pertenecen a la misma distribución aleatoria si los parámetros c no son distintos de cero. Las variables explicativas que usamos son el número de firmas, el número de electores inscritos en el REP para el momento del Reafirmazo, el número de nuevos electores (inscritos con posterioridad al Reafirmazo) y el número de electores que no votaron. Estimamos la ecuación en logaritmos.

Los resultados son muy claros, tal como lo indica el Cuadro 10. El término de interacción D * Firmas indica que la elasticidad de las firmas en votos es 10.5 por ciento más alta en los centros auditados que en los no auditados. Es decir, las firmas recolectadas en los centros auditados el 18 de agosto generan 10 por ciento más votos Sí que en el resto de los centros. El valor del estadístico de la t de Student es de 2.73. La probabilidad que esto sea por casualidad es menor al 1 por ciento (indicado por los 3 asteriscos en el Cuadro). El coeficiente sobre nuevos electores también es distinto con un nivel de confianza del 1 por ciento mientras que el coeficiente respecto de los electores no votantes es distinto con un nivel de confianza del 10 por ciento.

Para ilustrar lo inusual de este resultado construimos 1000 muestras aleatorias de 200 centros a partir del universo de centros no auditados. Estimamos la misma ecuación y calculamos el estadístico de t de Student para el término D * Firmas. El resultado se encuentra en el Cuadro 11. Como el Cuadro indica, un valor de dicho estadístico superior a 2.48 ocurre menos del 1 por ciento de las veces. En la muestra de la auditoría del 18 de agosto dicho valor es 2.73.

Cuadro 10. ¿Son los centros auditados representativos del universo de centros?

	Log SI
Log FIRMA	0.958
	(129.46)***
D * LFIRMA	0.105
	(2.73)***
Log Electores Reafirmazo	0.043
	(4.89)***
D * Log Electores Reafirmazo	-0.126
	(3.06)***
Log Electores Nuevos	0.595
	(23.64)***
D * Log Electores Nuevos	0.118
	(1.30)
Log Electores no votantes	-0.459
	(11.47)***
D * Log Electores no votantes	0.312
	(1.89)*
AUDIT	0.171
	(1.51)
Constant	0.254
	(9.14)***
Observations	4580
R-squared	0.97

Robust t statistics in parentheses
* significativo al 10%; ** significativo al 5%; *** significativo al 1%

Cuadro 11. Distribución de frecuencias del valor del estadístico de t de Student sobre el parámetro de firmas en 1000 regresiones estimadas en base a 1000 muestras extraídas aleatoriamente del universo de centros no auditados.

```
-------------------------------------------------------------
      Percentiles      Smallest
 1%     -2.60853      -3.342794
 5%    -1.832646      -3.233441
10%    -1.425525      -3.053542       Obs                1000
25%    -.8046502      -3.053519       Sum of Wgt.        1000

50%    -.0189599                      Mean          -.0191664
                        Largest       Std. Dev.      1.104314
75%     .7440667       3.232639
90%     1.360018       3.658616       Variance       1.219509
95%     1.770322       3.975739       Skewness       .0747199
99%      2.48632       4.010863       Kurtosis       3.049892
```

Concluimos que los datos indican que los centros auditados son estadísticamente distintos de los centros no auditados. Ello implica que no constituyen una muestra aleatoria del conjunto de centros. En dichos centros, las firmas se transforman en un número mayor de votos que en la totalidad de los centros. La probabilidad que esto sea una coincidencia es menor al 1 por ciento. Este resultado tiende a confirmar las dudas planteadas respecto a la confiabilidad de la auditoría.

Conclusiones

Este informe rechaza ciertas hipótesis de fraude pero no otras. No encontramos validez empírica para la muy discutida hipótesis de los topes numéricos. Tampoco encontramos sustento a cualquier hipótesis que implique alterar de manera diferencial las máquinas de un mismo centro. Una manipulación de este tipo alteraría las diferencias porcentuales en forma tal que hubiese violado la varianza esperada a nivel de centro y hubiese sido detectado por este análisis.

Toda hipótesis de fraude debe suponer una alteración similar en todas las máquinas de un mismo centro. Si esto hubiese sido realizado de manera homogénea en todos los centros de votación del país, los métodos utilizados en este estudio --y de hecho, probablemente ninguno de los métodos estadísticos--, podrían identificarlo. Lo que permite encontrar una prueba para examinar la posible existencia de fraude es precisamente el tratamiento heterogéneo de los distintos centros de votación.

Para realizar esta prueba utilizamos dos indicadores imperfectos, aleatorios e independientes de la intención de voto. Nuestra definición de fraude consiste en la existencia de una diferencia entre la intención de voto del elector y los votos registrados por el CNE. Nuestros dos indicadores, por más imperfectos que sean, están correlacionados con la intención del elector, pero no con el fraude. Si estos se utilizan independientemente en sendas regresiones para estimar la relación entre estos y los votos, el término de error o desviación reflejará no solo la imperfección del instrumento sino también el fraude. Si las dos desviaciones están correlacionadas, esto indica que hay un elemento común de desviación en ambas. Este elemento sería el fraude. Además, para ser consistente con la hipótesis de fraude, esta correlación debe ser positiva. En los centros donde un indicador proyecta equivocadamente más votos el otro también lo hace: ello indicaría que allí la incidencia de fraude es más alta.

Esto es precisamente lo que encontramos. Nuestros dos indicadores son el número de personas registradas en cada centro de votación que firmaron en el Reafirmazo de noviembre de 2003 y los exit polls realizados el 15 de agosto de 2004, día del Referéndum Revocatorio por Súmate y Primero Justicia. El resultado se mantiene si controlamos por los cambios en el registro electoral y en la tasa de participación. Además, el resultado es robusto a cambios en la forma funcional de la relación (lineal, logarítmica, porcentual). El resultado no se debe a efectos estadísticos espúreos (errores en las variables o la posible presencia de coeficientes aleatorios), pues se mantienen cuando corregimos por estos usando estimadores basados en variables instrumentales.

Como dijéramos anteriormente, cualquier hipótesis de fraude debe suponer que se alteraron en forma proporcional los resultados de todas las máquinas de un mismo centro. Esto supone necesariamente algún mecanismo de coordinación. En teoría este pudiera estar en el software o en la comunicación con el Centro de Totalización. Por estas razones es útil destacar los siguientes antecedentes:

- Las máquinas tenían la capacidad de comunicarse bi-direccionalmente con el Centro de Totalización y esta comunicación ocurrió.
- Las máquinas se comunicaron con el Centro de Totalización antes de imprimir las actas con los resultados, lo que abre la posibilidad de que hayan sido instruidas a imprimir resultados distintos de los reales
- No se permitió la entrada a testigos de la oposición o de los observadores internacionales en el Centro de Totalización durante la jornada electoral

El sistema de votación adoptado en Venezuela genera unas papeletas que son chequeadas por el votante y colocadas en cajas, las cuales están sujetas a ser auditadas de manera aleatoria. Un esquema de fraude debe contemplar como evitar la detección por la vía de la auditoría.

Una posibilidad es que los centros a ser alterados se escojan aleatoriamente. Ello genera dos tipos de centros de votación: aquellos que fueron alterados y los que no lo fueron. Salvo por este aspecto, cada grupo constituye una muestra representativa del país en términos regionales y sociales y por lo tanto, cada sub-muestra de estos también será una muestra representativa, salvo por el hecho de que unas fueron alteradas y otras no. Ahora bien, si se tiene control del programa que selecciona las cajas a ser abiertas en un proceso de auditoría, este puede escoger de manera aleatoria las cajas de aquellos centros que no fueron alterados y dicha muestra parecerá aleatoria en todos los aspectos salvo en lo que se refiere al fraude. En este sentido, cabe destacar como antecedente que el CNE se negó a utilizar el programa generador de números aleatorios ofrecido por el Centro Carter para la auditoría del 18 de agosto y en vez se utilizó el programa del CNE instalado en su propio computador.

Nuestro análisis indica que la muestra seleccionada para realizar la auditoría del 18 de agosto de 2004 no es aleatoria y representativa del conjunto de centros de votación. En dicha muestra, la elasticidad de las firmas frente a los votos es 10 por ciento más alta y la posibilidad de que esto sea aleatorio es significativamente menor al 1 por ciento. Repetimos nuestro análisis escogiendo aleatoriamente 1000 muestras de centros no auditados y este resultado ocurre en menos del 1 por ciento ellas.

En conclusión, este estudio rechaza ciertas hipótesis de fraude, pero indica otras que son compatibles con los datos estadísticos.

En estadística es imposible confirmar una hipótesis, pero si es posible rechazarla. Como dijera Karl Popper, el observar 1000 cisnes blancos no demuestra la veracidad de la tesis de que todos los cisnes son blancos. Sin embargo, observar un cisne negro si permite

rechazarla. Parafraseando a Popper, nuestro cisne blanco es que no hubo fraude. Los resultados que obtenemos constituyen un cisne negro. La hipótesis alternativa de que sí hubo fraude es consistente con nuestros resultados y por tanto no podemos rechazarla.

Appendix 1. Identifying fraud using imperfect correlates of the intention to vote

Assume that the fraud is defined as the difference between the votes for SI actually collected and an unobservable variable that is the intention of voting of the voters that showed up. We define the first one as Vi, the intention of voters as Xi, and the fraud as Fi.

Vi = Xi + Fi

There are also two additional measures of the intention of voters: the exit polls (Ei) and the signatures (Si) in each of the centers. These measures, however, are imperfect. We assume a very general form of that imperfection – a random coefficient model. However, to make the point clear we start with a simpler form of errors and then generalize them. Assume that

Ei = a*Xi + epsi
Si = b*Xi + etai

Where we are assuming that the exit polls are possibly a biased estimate of the intention to vote: a can be smaller than one. The signatures (Si's), as well, could be a biased measure. Both equations have an error (epsi and etai) that take into account the fact that both the exit polls and the signatures are very imperfect measures of the voter's intentions – even the biased measured intentions. We assume that these errors are uncorrelated among themselves and with the fraud.[3]

How can we detect the fraud? The fraud can only affect the actual votes, not the exit polls, nor the signatures. In other words, the fraud is a displacement of the distribution of votes that is not present in the other two measures. Statistically, this means that the fraud could be detected by using the exit polls and the signatures as predictors of the voting process and analyzing the correlation structure of the residuals. Under the assumption that all residuals are uncorrelated – which makes sense given the definitions we have adopted – then the correlation of residuals is an indication of the magnitude of the fraud.

The particular procedure used to detect the fraud is the following:

1. Estimate the regression of Vi on Ei plus controls and recover the residual. This residual has two components: the fraud and the errors in variables residual due to the fact that the exit polls are noisy.

2. Estimate the regression of Vi on Si plus controls and recover the residual. This residual has two components: the fraud and the errors in variables residual due to the fact that the signatures are an imperfect measure of the intention of voters.

[3] This is a reasonable assumption considering that the signatures were collected at different times and conditions than the exit polls.

Notice that these two residuals are correlated. First, because both have the fraud as an unobservable component, and second, because the right hand side variables are correlated and there is errors in variables in the regression.

3. Estimate the regression of Vi on Ei plus controls using Si as an instrument. Recover the residual. Notice that in our model, because etai is uncorrelated with epsi and Fi, we can use Si as an instrument to correct for the error in variables.

4. Using the same logic estimate Vi on Si plus controls, and using Ei as the instrument. Recover the residual.

   In this case, because the two coefficients are supposed to have solved the problem of error in variables the residuals can only be correlated if there is a common component – which in our case is the definition of the fraud.

This procedure actually detects how important the fraud is. The section that follows first explains why this procedure indeed is able to identify the fraud. After that we also analyze the possibility that the fraud is correlated with the signatures – which is likely given what we have argued about the stochastic properties of the votes per machine and center. Finally, we present evidence.

Procedure

This section describes the procedure in a simple model

OLS estimation with no correlation between fraud and intention to vote

Running the OLS regression of Votes on Exit Poll is:

Vi = Xi + Fi
Ei = a*Xi + epsi

Where

Xi = (1/a)*Ei - (1/a)*epsi

Subtituting in the voting equation

Vi = c1*Ei + psi1

Where

psi1 = Fi – (1/a)*epsi

In this model, the OLS coefficient is

c1ols = a*var(Xi) / (a^2 * var(Xi) + var(eps))

which is always smaller than 1/a which is the true coefficient. This means that the residual from the regression (psi1) is

psi1 = Fi +(1/a-c1ols)*Ei – (1/a)*epsi

We can do the same thing for the signatures. Notice that everything is symmetric so the equations are almost identical.

Vi = Xi + Fi
Si = b*Xi + etai

Which means that

Xi = (1/b)*Si - (1/b)*etai

Substituting Xi in the Vi equation

Vi = c2*Si + psi2

Where

psi2 = Fi - (1/b)*etai

In this model, the OLS coefficient is

c2ols = b*var(Xi) / (b^2 * var(Xi) + var(eta))

which is always smaller than 1/b – it is only equal to 1/b when the variance of eta is zero. The outcome is that the residual will be

psi2 = Fi +(1/b-c2ols)*Si – (1/b)*etai

Notice that the two residuals are correlated. Under the assumption that epsi and etai are uncorrelated, and also uncorrelated with the fraud there are two components that create the correlation among these residuals: the fraud, and the errors-in-variable bias.

cov(psi1,psi2) = var(Fi) + (1/a-c1ols)*(1/b-c2ols)*cov(Ei,Si)

The first term is the variance coming from the fraud, while the second term comes from the variance due to the error-in-variables that is present in both Ei and Si. Notice that we are assuming that the errors in variables are independent. The covariance arises because the error-in-variables downward biases both coefficients ($c1ols<1/a$ and $c2ols<1/b$) and because the exit polls and the signatures are correlated.

Instrumental Variables with no correlation between fraud and intention to vote

Under our assumptions, we have an easy solution to the error in variables in both regressions. Notice that etai and epsi are uncorrelated and that etai is uncorrelated with Fi. Additionally, Ei and Si are correlated because both measure the same factor (Xi). This means that Si can be used for instrumenting Ei and Ei for instrumenting Si. The outcome is as follows:

$$Vi = c1*Ei + Fi - epsi$$

The IV estimate is

$$c1iv = cov(Si'Vi) / cov(Si'Ei)$$
$$c1iv = b*var(Xi) / a*b*var(Xi)$$
$$c1iv = 1/a$$

which means that the residual is

$$psi1 = Fi - (1/a)*epsi$$

Notice that now the errors-in-variable component has disappeared. Similarly, if we run the regression for votes on signatures and using the exit polls as instrument we find:

$$Vi = c2*Si + Fi - (1/b)*etai$$

The IV estimate is

$$c2iv = cov(Ei'Vi) / cov(Ei'Si)$$
$$c2iv = var(Xi) / b*var(Xi)$$
$$c2iv = 1/b$$

which means that the residual is

$$psi2 = Fi - (1/b)*etai$$

The correlation between the residuals of the two IV regression is

$$cov(psi1,psi2) = var(Fi)$$

So, a simple test is to compare these two covariances, and determine if they are statistically different. Furthermore, if the covariance of the IV residuals is different from zero, then we have an estimate of the importance of the fraud.

Correlated Fraud

OLS estimation with correlation between fraud and intention to vote

The previous exercise has assumed that the fraud is uncorrelated with the signatures, but as we have argued in the previous section, this is unlikely. In fact, most probably, the fraud is correlated with the signatures. Let us repeat the previous exercise allowing for any covariance structure between the fraud and the signatures. Running the OLS regression of Votes on Exit Poll we obtain the same result as before:

Vi = Xi + Fi + f*Si
Ei = a*Xi + epsi

Where the residual of the voting equation has the independent term Fi and the part of the fraud that is correlated with the signatures (f*Si). In this model, the OLS coefficient is the same as before

c1ols = cov(Vi,Ei)/Var(Ei)
= (a*var(Xi) + a*f*cov(Xi,Si))/ (a^2 * var(Xi) + var(eps))
= (a*var(Xi) + a*f*b*var(Xi))/ (a^2 * var(Xi) + var(eps))

which now we can’t be sure is smaller than 1/a as before. This depends entirely on the sign of f. however, if f is negative (as we will mostly argue in this paper), then the bias downward is stronger than in the pure case. This means that the residual from the regression (psi1) is

psi1 = Fi +(1/a-c1ols)*Ei – (1/a)*epsi+f*Si

For the signatures model

Vi = c2*Si + Fi - (1/b)*etai + f*Si

Where all the three terms in the right hand side are part of the residual. The OLS coefficient is

c2ols = cov(Vi,Si)/Var(Si)
= (b*var(Xi) + f*var(Si))/ (b^2 * var(Xi) + var(eta))

where the additional term in the numerator is coming from the correlation structure between the signature and the fraud. A plausible assumption is that the fraud is usually a negative variable (in our specification) which we could expect to be larger the larger the

signature is. This means that the coefficient, f, is likely to be negative. This means that the bias in c2ols is downward and stronger than just with errors in variables.

The residual is

psi2 = Fi +(1/b-c2ols+f)*Si – (1/b)*etai

Notice that the two residuals are correlated as before, but now there are two additional terms.

cov(psi1,psi2) = var(Fi) + (1/a-c1ols)*(1/b-c2ols)*cov(Ei,Si)
+ (1/a-c1ols)*f*cov(Ei,Si) + f^2*var(Si)

Instrumental variables estimation with correlation between fraud and intention to vote

Lets see what is the implication for this correlation when we do the instrumental variables approach. For the first equation we have:

Vi = c1*Ei + Fi – (1/a)*epsi+f*Si

The IV estimate is

c1iv = cov(Si'Vi) / cov(Si'Ei)
c1iv = (b*var(Xi) +f*var(Si))/ b*a*var(Xi)

c1iv = 1/a+(f/ba)*var(Si)/var(Xi)

which means that the residual of the IV regression is

psi1 = Fi – (1/a)*epsi + f*Si + (1/a-c1iv)*Ei

It is easy to show that c1iv is closer to 1/a than c1ols, which means that the residual, psi1, has a component coming from the error in variables that is smaller than from the OLS regression.

The IV regression of the Si specification is as follows:

Vi = c2*Si + Fi - (1/b)*etai + f*Si

The IV estimate is

c2iv = cov(Ei'Vi) / cov(Ei'Si)
c2iv = a*var(Xi) / a*b*var(Xi)
c2iv = 1/b

Notice that in this case the estimate of the IV in the voting on signatures is exactly the true coefficient. Exactly what we obtained in the previous section. Why? Simply because the exit polls do not have the error component coming from the fraud. The fraud – which is the residual in the voting equation – cannot be correlated with the exit polls or its innovations. Signatures, on the other hand, are correlated with the fraud. This is indeed the assumption how the fraud was performed.

This makes exit polls a good instrument for signatures, but signatures is not a good instrument for exit poll. The residual in the second IV regression is:

psi2 = Fi - (1/b)*etai + f*Si

Comparison of the covariance

Lets compare the two covariances: the covariance for the OLS residuals with the covariance with the IV residuals. The OLS residual have

cov(psi1,psi2)_OLS = var(Fi) + (1/a-c1ols)*(1/b-c2ols)*cov(Ei,Si)
+ (1/a-c1ols)*f*cov(Ei,Si) + f^2*var(Si)

while the covariance for the IV estimates is

psi1 = Fi – (1/a)*epsi + f*Si + (1/a-c1iv)*Ei
psi2 = Fi - (1/b)*etai + f*Si

cov(psi1,psi2)_IV = var(Fi) + f*(1/a-c1iv)cov(Ei,Si) + f^2*var(Si)

First, notice that as before, if there is no fraud the covariance of the IV residuals should be zero. Furthermore, this last covariance reflects different forms of fraud. If the fraud is a random variable shifting the distribution (or equivalently that f=0 and Fi<0) the covariance is the same as before:

cov(psi1,psi2)_IV = = var(Fi)

if the fraud is not introduced as a random variable but as a shift in the distribution correlated with the signatures (f<0 and Fi=0) then the covariance of the IV residuals is

cov(psi1,psi2)_IV = f(1-c1iv)cov(Ei,Si) + f^2*var(Si)

Only if Fi=0 and f=0 will produce a zero covariance of the IV residuals. In reality, if there is a fraud, probably both aspects will enter and the covariance is a combination of the two.

The next question is, what is the direction of the change in the covariance, from OLS to IV?

$$\text{cov_OLS} - \text{cov_IV} = (1/a-\text{c1ols})*(1/b-\text{c2ols})*\text{cov}(Ei,Si) + (\text{c1iv}-\text{c1ols})*f*\text{cov}(Ei,Si)$$

where the two terms are easily signed. Lets start with the first term. We know that the error in variables together with a negative f implies that both OLS estimates are downward biased. We also know that a reasonable set of assumptions imply that signatures and exit polls are positively correlated. Hence, the first terms is a multiplication of three positive elements. Let us turn our attention now to the second term. We know that c1iv is closer to 1/a than c1ols. This means that the term in brackets is negative, and we have been analyzing only the case in which f is negative. Hence, the covariance of the OLS residuals has to be larger than the covariance of the IV residuals.

Notice that if f were positive we could not have made this claim. And there would be circumstances in which the covariance actually goes up after instrumenting.

In the empirical section we implement this strategy by calculating the covariance both using OLS and instrumental variables and observe a reduction in the covariance, as would be consistent with the presence of f being negative (i.e. against the Yes vote). In addition, we find that cov_IV is positive and statistically significant which is informative of the fact that f is significantly different from zero.

Appendix 2. Identifying whether the audited center were chosen randomly

Sophisticated frauds are hard to detect. According to Rubin if all the machines are affected using the same procedure then the fraud is statistically undetectable. The advantage of the Venezuelan case is that there was the possibility of audits and therefore, some of the CDV's were required to be untouched. Because the government was the one that would choose which centers would be audited, it is imaginable that they knew which centers were not touched and lead the auditors toward those centers.

First, lets discuss the form of the fraud (the theory). Then, we provide the evidence.

The theoretical idea of a voting process is that it is an imperfect but unbiased measure of the popular intention

$Vi = Xi + Fi$

Where Xi is the voting intention in each CDV, Vi are the total votes, "i" indicates the CDV, and Fi is the fraud. It is important to indicate that if Fi is a shift with the same distribution as Xi (meaning both are normally distributed) then Vi is the sum of two normal distributions, which is also a normal distribution and it would be hard to identify any anomalies in the data. The fraud would be undetectable.

Under this assumption, in the case of Venezuela, the fraud could not be normally distributes because the system had to allow some centers to be untouched so that they could be audited. Interestingly, the government did not allow the Carter Center to choose the voting centers randomly. It was the government the one that assigned the voting centers. So, if there is a fraud, it is imaginable that we could detect the shift in the population by comparing the audited population with the non-audited population.

We have a variable that is correlated with the intention of voters – the signatures. We assume that the signatures follow

$Si = b(i)*Xi + etai.$

where Si are the number of signatures, b(i) is a random coefficient mapping the intention of voters to the total number of signatures, and etai is a random disturbance indicating the noise involved in the measurement of the intentions.

This specification allows the signatures to be a biased estimator of the voters intentions (b(i) on average could be less or larger than one).

Additionally, Fi could be correlated or not with the number of signatures. We will adopt, then the following specification: Fi + f(i)*Si to encompass both possibilities.

What is the OLS estimate of Vi on Si?

$$Vi = c*Si + psi$$

Notice that the reduced form model is the following:

$$Vi = Xi + Fi + f(i)*b(i)*Xi + f(i)*etai$$
$$Si = b(i)*Xi + etai.$$

The OLS coefficient is the covariance between these two variables divided by the variance of the signatures. These variances include not only the variances of the shocks but also the variances of the coefficients. For notational convenience assume

$$b(i) = b + bi$$
$$f(i) = f + fi$$

where bi and fi conditional on Xi have mean zero and finite variance. Under these assumptions the OLS coefficient is

$$cols = \left(b(1+fb)*var(Xi) + f*var(eta)\right) / \left(b\text{^}2*var(Xi)+var(eta)\right)$$

which obviously is different from 1/b – which is the limit if f=0 and var(eta)=0. There are several biases in this coefficient worth highlighting: first the error-in-variables which is the result of var(eta) being different from zero. This bias is the attenuation bias and reduces the coefficient. So $cols<1/b$. Second, there is the bias introduced by the fraud component that is correlated with the signatures (f). we will assume throughout the paper that f is negative, if that is the case, notice that the denominator is reduced by this bias, which means that cols is farther from 1/b. In summary, both biases are working in the same direction.

Let us see how the residuals look

$$psi = Fi + (1+(f(i)-cols)*b(i))*Xi + (f(i)-cols)*etai$$

Notice that even if the structural shocks are homoskedastic (var(Xi), var(eta) and the variances of the random coefficients) the variance of the residuals is going to be heteroskedastic. There is a term multiplying the residuals that is related to the random coefficient model.

Therefore, it is not surprising that indeed the standard deviation of the residuals of this regression at the parroquia level are correlated with the predicted residuals of the regression. Part of that correlation is coming from the fraud (obviously), but also part of that correlation is the result of the random coefficient model. Let us see.

If f(i) and Fi are zero, then

$$psi = (1-cols*b(i))*Xi -cols*etai$$

$cols = \left(b*var(Xi) \right) / \left(b^2*var(Xi)+var(eta) \right)$

notice that unambiguously cols is smaller than 1/b which means that the expected value of cols*b(i) is smaller than E(b(i)/b)=1. Therefore the term in the first bracket of the residuals has a posi tive expected value and the variance of psi will depend on Xi. This source of correlation is not interesting (in terms of fraud) and therefore, we require a better test to differentiate between a random coefficient model and one with fraud.

The idea is to compare a "supposedly random" sample with the total sample. If the sample is truly a random coefficient model, and the innovations to the coefficients (bi and fi) are truly orthogonal to the other shocks, then any sub-sample – any sub-sample – should have the same properties as the full sample.

This is exactly what is done when agencies collect surveys on consumption, industry production, etc. If the sub-sample is a random draw it is representative of the population, or full-sample. For instance, assume that we are interested in studying consumption patterns. We know that consumption depends on the level of income, education, race, gender, age, religion, etc. Furthermore, there is no particular reason why we have to assume that a one percent increase in income will imply the same increase of consumption to all the individuals in the sample. In other words, it is reasonable to assume that the coefficients from income to consumption are random. However, if the model is well specified (meaning that all the controls that have to be in the right hand side are there), then the coefficients are truly random and independent of everything else. If we pick a random sub-sample of the population – a representative sample – the behavior of those individuals is a good proxy of the behavior of the population. This is standard in all micro data models where always we make inference about the population by looking at a smaller sample. This is exactly what we do here.

Any sub-sample of centers should have the same behavior as the whole. This does not mean that the coefficients estimated are going to be the same. What it does mean is that the differences cannot be statistically significant. We can, for example, choose as the random sub-sample, the centers that the government allowed the Carter Center to audit. Why is this a good sample? Well, because we know that a shift of the full distribution is statistically undetectable, the fraud cannot only be found if we concentrate in the sub-sample that, ex-ante, has a lower likelihood of being tainted.

Therefore, we split the sample between those that were audited and those that were not audited. As is discussed in the paper, we find them to be statistically different.

Differences in the samples in terms of their OLS estimates

As was mentioned above, the fraud introduces a bias in the regression coefficient if it is correlated with the signatures in the centers. To clarify the exposition we show the OLS coefficient of the bivariate model with and without fraud:

$$cols_fraud = \left(b(1+fb)*var(Xi) + f*var(eta)\right) / \left(b^2*var(Xi)+var(eta)\right)$$

$$cols_Nofraud = \left(b*var(Xi)\right) / \left(b^2*var(Xi)+var(eta)\right)$$

which implies that

$$cols_fraud = cols_Nofraud + f*\left(b*var(Xi) + var(eta)\right) / \left(b^2*var(Xi)+var(eta)\right)$$

which under our assumptions that $f<0$ implies that the OLS coefficient of the fraud sample is smaller than the OLS coefficient of the no-fraud sample. Additionally, it should be the case that these two coefficients are statistically different from zero, because otherwise the changes in the coefficients are mainly explained by the small sample properties of OLS and not by fraud.

To test for this possibility we estimate our preferred estimation allowing for interactions of all the right hand side variables with a dummy that takes value of one when the center was one of the ones assigned to be audited. The regression is

$$Vi = c2*Si + c3*Si*D + c4*NewElectors + c5*NewElectors*D + c6*Participation + c7*Participation*D + c0 + c1*D$$

Where we are predicting the total votes by the signatures (Si) and the signatures interacted with the dummy for audited centros (D). We introduce several controls, obviously allowing for different constant terms in the two sub-samples (c0 + c1*D) and controlling for the increase in the universe of voters (NewElectors) and for the participation in the voting center (Participation).

The coefficients of interest are c3. If there is fraud (and hence there is a shift in the distribution), then c3 should be positive and statistically different from zero. Why? As was shown before, under the assumption that the fraud reduced the number of votes for “Si” the OLS coefficient in the full sample (c2) is smaller than the true one (c2+c3). Notice that this is what we find in the empirical results.